An investigation into the effectiveness of simulation-extrapolation for correcting measurement error-induced bias in multilevel models

Christopher Custer

April 17, 2015

Abstract

This paper is an investigation into correcting the bias introduced by measurement errors into multilevel models. The proposed method for this correction is simulation-extrapolation (SIMEX). The paper begins with a detailed discussion of measurement error and its effects on parameter estimation. We then describe the simulation-extrapolation method and how it corrects for the bias introduced by the measurement error. Multilevel models and their corresponding parameters are also defined before performing a simulation. The simulation involves estimating the multilevel model parameters using our true explanatory variables, the observed measurement error variables, and two different SIMEX techniques. The estimates obtained from our true explanatory values were used as a baseline for comparing the effectiveness of the SIMEX method for correcting bias. From these results, we were able to determine that the SIMEX was very effective in correcting the bias in estimates of the fixed effects parameters and often provided estimates that were not significantly different than those from the estimates derived using the true explanatory variables. The simulation also suggested that the SIMEX approach was effective in correcting bias for the

random slope variance estimates, but not for the random intercept variance estimates. Using the simulation results as a guideline, we then applied the SIMEX approach to an orthodontics dataset to illustrate the application of SIMEX to real data.

1 Introduction

Measurement error problems arise when certain variables in a statistical analysis are measured inaccurately. The true value, X_t, is often unobserved directly and is instead observed with additional error (Fuller, 1987). This is a common and problematic source of bias in statistical models, resulting in incorrect inferences which can be very costly for the researcher. For example, medical and epidemiological data will often contain measurement error and the inferences obtained from these models are critical to the research being performed. Measurements such as blood pressure and vitamin levels, among many others, are often measured with error and this may cause misleading and incorrect results for the researcher's analysis. The inferences made from a study of this nature are obviously incredibly important and need to be as accurate and reliable as possible. From this, there is a need to develop tools and methods that can overcome the bias introduced by measurement error.

In general, statistical estimators are desired to contain certain properties, such as being unbiased, consistent and minimal variance. In regression models, for example, parameter estimates are computed under the assumption that explanatory variables are measured with exactness. When these variables are measured with error, these estimators often lose their desirable qualities. In many situations, it is impossible to eliminate measurement error and so we need to establish estimators that preserve the desired properties, despite the additional error.

This project investigates the effects of measurement error in certain covariates in multilevel models. Multilevel models are regression models that contained both fixed and random effects.

We investigate how measurement error in level 1 covariates affect the estimation of these fixed and random effects. In addition, we investigate the use of a method known as simulation-extrapolation (SIMEX) to correct the bias introduced by measurement error (Cook and Stefanski, 1994). SIMEX is a general and widely used method in measurement error modeling. SIMEX has been shown to be effective at correcting bias in a variety of regression models. We have not seen an example of its application to multilevel models and so our goal is to investigate the effectiveness of SIMEX for parameter estimation in multilevel models.

This paper will first provide a more in-depth background of measurement error, SIMEX, and multilevel models. We describe and report the results of a simulation designed to investigate the effects of measurement error and the SIMEX method in multilevel models. Finally, we apply SIMEX to an actual data set as an example of its real world applicability.

2 Background

2.1 Measurement error

Measurement error is a common source of bias in many statistical analyses, including regression analysis. In regression models, measurement error in covariates can cause bias in estimated model parameters, and can lead to incorrect inferences. Since the focus of this study is the effect of measurement error on parameter estimation in multilevel models, it is important to first define the general effects of measurement error.

In the classical measurement error model, we assume that a variable of interest, X, is measured with independent, additive error having mean 0. That is, X is measured as W where

$$W = X + U, \tag{1}$$

$E(U) = 0$ and U is independent of X. The additional assumption that the measurement error is normally distributed, $U \sim N(0, \sigma_u^2)$, is also commonly made. As will be the case in our study, it is also common, and often required, that σ_u^2 is known or accurately estimated when working with measurement error models.

To examine the impact that measurement error has on estimation of parameters, we will look at a few examples. Suppose the random sample $X_1, \ldots, X_n$ is observed as $W_1, \ldots, W_n$ according to the classical measurement error model in equation (1). First consider the effect of measurement error on the estimation of a population mean, μ_x. In the absence of measurement error the usual sample mean $\overline{X}$ is unbiased for μ_x and has variance $V(\overline{X}) = \sigma_x^2/n$. Replacing the true measurements with the error-prone measurements yields the estimator $\overline{W}$, and it is straightforward to show that

$$E(\overline{W}) = \mu_x \qquad \text{and} \qquad V(\overline{W}) = \frac{\sigma_x^2 + \sigma_u^2}{n}.$$

This implies that while our sample mean remains an unbiased estimator, its variability is inflated, and increases as the variance of the measurement error, σ_u^2, increases.

Next consider estimating the population variance, σ_x^2. In the absence of measurement error, the usual sample variance,

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ_x^2. When X is observed only as W, the sample variance becomes

$$\begin{aligned} s_W^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \bar{X}) + (U_i - \bar{U})]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \bar{X})^2 + 2(U_i - \bar{U})(X_i - \bar{X}) + (U_i - \bar{U})]^2 \\ &= s_X^2 + 2s_{XU} + s_U^2 \end{aligned}$$

and therefore

$$\begin{aligned}E(s_W^2) &= E(s_X^2 + 2s_{XU} + s_U^2) \\ &= E(s_X^2) + 2E(s_{XU}) + E(s_U^2) \\ &= \sigma_X^2 + \sigma_U^2\end{aligned}$$

Thus through the introduction of measurement error, we see that the sample variance is no longer unbiased, and this bias increases as the measurement error variance increases. It should also be noticed that given this result, if the measurement error variance is known, it would be simple to construct an unbiased estimator, such as $\widehat{\sigma}_X^2 = s_W^2 - \sigma_U^2$. It is easy to see that $E(\widehat{\sigma}_X^2) = E(s_W^2 - \sigma_U^2) = \sigma_X^2$.

Finally, consider estimation of the slope parameter in the simple linear regression model. Here we assume that the variable X is related to a response Y according to the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the regression errors ϵ_i are independent with mean 0 and constant variance. Recall that in the absence of measurement error, β_1 is estimated by $\widehat{\beta_1} = s_{xy}/s_x^2$. Under very general assumptions this is a consistent (even unbiased) estimator of the true slope. That is,

$$\widehat{\beta_1} = \frac{s_{xy}}{s_x^2} \xrightarrow{P} \frac{\sigma_{xy}}{\sigma_x^2} = \beta_1$$

Again suppose that X is measured according to the classical measurement error model, and further assume that U and ϵ are independent. Replacing X with W, the estimate slope becomes

$$\widehat{\beta}_{1,w} = \frac{s_{wy}}{s_w^2} = \frac{s_{xy} + s_{uy}}{s_w^2}.$$

Note that $s_{xy} \xrightarrow{P} \sigma_{xy}$, $s_{uy} \xrightarrow{P} \sigma_{uy} = 0$ and $s_w^2 \xrightarrow{P} \sigma_x^2 + \sigma_u^2$. Thus

$$\widehat{\beta}_{1,w} \xrightarrow{P} \frac{\sigma_{xy}}{\sigma_x^2 + \sigma_u^2} = \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \right)$$

This effect is known as attenuation, causing the slope to approach zero as the measurement error increases (Fuller 1987). From above, we can see again that as the measurement error variance increases, the bias increases and the estimated slope approaches zero. On the other hand, as the measurement error variance approaches zero, the bias approaches zero, as well.

In the case of known or estimable measurement error variance, a consistent estimator of β_1 that relies only on the observable data is constructed as

$$\widehat{\beta}^*_{1,w} = \widehat{\beta}_{1,w}\left(\frac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2}\right)$$

These simple examples illustrate some key concepts in measurement error modeling. First and foremost is that the measurement error creates bias in most of our estimates. This bias is directly related to the amount of variability there is within the measurement error. As the measurement error increases, the bias will increase. However, if the measurement error variance is known, this bias can be corrected using the methods described above, allowing for consistent and unbiased estimators using the observed data, W.

2.2 SIMEX

Simulation-extrapolation (SIMEX) is a simulation-based approach for removing the measurement error bias in parametric models (Cook and Stefanski 1994). The method is very general and can be applied in a wide variety of statistical analyses including regression modeling. It is a relatively simple method but requires that the measurement error variance is known or well-estimated. The method uses the fact illustrated above that as measurement error increases, the bias in the estimated parameters increases. The basic idea of the method is to first add simulated measurement errors with increasing variance to the data, estimate the parameter of interest, and determine a trend between the resulting estimates and the measurement error variance. The SIMEX estimate of the parameter is the extrapolation of this trend to a point representing a measurement error variance of

0, corresponding to the estimator computed from data with no measurement error, denoted $\widehat{\theta}_{true}$. Cook and Stefanski (1994) show that this method is equivalent, or asymptotically equivalent, to method of moments estimation.

To understand the SIMEX process, first define

$$W_j = X_j + U_j$$

where W_j is our observed variable, X_j is our true variable, and U_j is our measurement error, independent of X_j, with $U_j \sim N(0, \sigma_u^2)$ where σ_u^2 is known. Next, let $\lambda > 0$ and $Z_j \sim N(0,1)$. Z_j is our additional known measurement error, or pseudo-error, necessary to the SIMEX estimation process. Now define

$$\begin{aligned} W_j(\lambda) &= W_j + \sigma_u \sqrt{\lambda} Z_j \\ &= X_j + U_j + \sigma_u \sqrt{\lambda} Z_j \\ &= X_j + U_j^* \end{aligned}$$

and thus $U_j^* \sim N(0, \sigma_u^2 + \lambda\sigma_u^2)$. From this, it is easy to see that $\mathrm{Var}(U_j^*) = (1+\lambda)\sigma_u^2$ and so $W_j(\lambda)$ represents an observation to which additional measurement error is added. If we let

$$\widehat{\theta}(\lambda) = g(W_1(\lambda), ..., W_n(\lambda))$$

then $\widehat{\theta}(\lambda)$ represents our parameter estimate for the given value of λ. It follows that

$$\hat{\theta}(-1) \approx \hat{\theta}_{true}$$

In simpler terms, by setting $\lambda = -1$, we are forcing $U_j^* \sim N(0,0) = 0$ and thus $W_j = X_j$, the true value of our explanatory variable.

In order to implement the SIMEX method, the following steps should be taken:

1. Choose severals values of λ such that $0 < \lambda_1 < ... < \lambda_m$

2. Using a simulation, estimate $\widehat{\theta}(\lambda_k)$ for each $k = 1, ..., m$

3. Plot $\widehat{\theta}(\lambda_k)$ vs λ_k and extrapolate to $\lambda = -1$ (see Figure 1)

There are many ways to accomplish step 2, these steps depending on the structure of your SIMEX estimates. One key factor in the process of estimating $\hat{\theta}(\lambda_k)$ is that it requires many estimated simulations for each value of λ_k, from which the mean of those simulated values is used in the extrapolation process. Cook and Stefanski (1994) suggest using Monte Carlo simulation for the estimation process described in step 2. Specifically for each λ_k, a large number B of Monte Carlo replicates are repeated, each resulting in the estimate $\widehat{\theta}_i(\lambda_k)$. Then the Monte Carlo estimator of $\theta(\lambda_k)$ is

$$\widehat{\theta}(\lambda_k) = \frac{1}{B} \sum \widehat{\theta}_i(\lambda_k) \tag{2}$$

For the extrapolation in step 3, they recommend using one of three methods to establish the trend between λ_k and $\hat{\theta}(\lambda_k)$. The first, which is shown as an example in Figure 1, is to fit a linear regression to describe the trend. The second method is to use a quadratic regression for cases where the trend is curved. Their third method is for when the trend appears to be nonlinear, and follows the form of $\mu(\lambda) = a + b/(c + \lambda)$. The authors of this method also note that when the meaurement error is normally distributed, each of the extrapolants is exact for certain estimators. However, they claim that even more important than the exactness of the estimate is the fact that at least one of these extrapolation methods will provide a sufficiently good estimate.

2.3 Multilevel Models

Multilevel models are extensions of the simple linear regression model that treat model coefficients as random variables instead of fixed effects. These models, sometimes referred to as mixed-effect or hierarchical models, are used when modeling the coefficients of a population as fixed effects is insufficient. A good example is one used by Goldstein (1995), which describes a study of math scores at age 8 and age 11 from children of different schools across London. These data could be used to create a simple linear regression to predict math scores at age 11 based on age 8 scores. However, it is also very likely that the trend between scores differs tremendously among schools. A multilevel model could be used to allow model parameters describing the relationship between age 8 and age 11 scores to vary randomly among schools. Such a model would allow estimation of a general, overall trend as well as the variability of the trend among students at different schools. For example, the estimates from this model could suggest some trend between the two age scores, but that the data also shows a high variability for trends among schools. Technically, predicted values could be formulated from this model, but the variance between schools would leave such estimates highly suspect.

For this study we will focus on multilevel models with a single explanatory variable having a linear relationship with a response. There are three ways to parameterize such models, and those are referred to as the random intercept, the random slope, and the random slope and intercept models. These are defined as their name implies, with an intercept that has a random component, a slope that has a random component, and a model where both slope and intercept contain random components. For example, the random slope and intercept model is defined as

$$Y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + e_{ij}$$

where the random components are typically defined as $b_{0j} \sim N(0, \sigma^2_{b0})$ and $b_{1j} \sim N(0, \sigma^2_{b1})$. Note that b_{0j} and b_{1j} are not necessarily independent of one another, thus we define $cov(b_{0j}, b_{1j}) = \sigma_{b01}$. Let $i = 1, 2, ..., m$ be defined as the number of level 1 units, and $j = 1, 2, ..., n$ be defined as the number of level 2 units. Level 1 units are defined as those clustered within each level 2 unit. For our school example, level 1 units would be each student and level 2 units would be each school (Goldstein 1999). If the random components are defined as above, then the exected value of Y is $E(Y) = \beta_0 + \beta_1 X$. This represents the average slope and intercept for the population. The random components, b_0 and b_1, describe the variability of the intercepts and slopes across level 2 units, respectively, around the population averages.

To relate this back to the math scores example, using a random slope and intercept model suggests that there is variability from the population in both the intercept and slope for each school. The fixed effects returned from the fitted model would represent the estimated intercept and slope for the population of schools sampled. The random components would represent how much each school's intercept and slope vary from the population. If the schools' intercept and slope tend to show large deviations from the population, then we would expect the fitted model to return large variance estimates for our random effects. The fitted model will also estimate the covariance between the random components. This would allow us to interpret the relation between our random variances. A large positive covariance would suggest that if we see an increase in the variability of our intercepts, we should expect to see an increase in the variability of our slopes. Just as in simple linear regression, measurement error is problematic in multilevel modeling (Goldstein 1995).

3 Simulation

We performed a simulation study to determine the effectiveness of the SIMEX method for correcting covariate measurement error-induced bias in parameter estimation for multilevel models. The main objective of our simulation study was to understand if and when the SIMEX method produces improved parameter estimates in multilevel models, compared to estimates that ignore measurement error in the covariate.

Data for simulations were generated as follows. First true values of the covariate X_{ij} were set and used to generate responses, Y_{ij}, according to a multilevel model. Next, measured values of the covariates, W_{ij}, were generated according to the classical measurement error model. Specifically,

$$Y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})X_{ij} + e_{ij} \quad (3)$$

$$W_{ij} = X_{ij} + U_{ij} \quad (4)$$

where $e_{ij} \sim N(0, \sigma_e^2)$, $b_{0j} \sim N(0, \sigma_{b0}^2)$, and $b_{1j} \sim N(0, \sigma_{b1}^2)$. The parameters of interest to our study are the fixed intercepts and slope, β_0 and β_1, and the corresponding random components' variance, σ_{b0}^2 and σ_{b1}^2. The values of these parameters were held constant for all simulated datasets, with $\beta_0 = 2, \beta_1 = 3, \sigma_{b0}^2 = 16, \sigma_{b1}^2 = 25$, and $\sigma_e^2 = 6.25$. The true values of the explanatory variable, X, were defined simply as a sequence of 1 to m, where m was defined as the number of level 1 units. This simplistic approach was used to limit the variation in the true values so that the effects of the measurement error variance would be more easily observed.

Four sets of parameter estimates were computed for each simulated data set. These were the true estimates, naive estimates, SIMEX using linear extrapolation (linear SIMEX), and SIMEX using quadratic extrapolation (quadratic SIMEX). The true estimates were estimated model parameters calculated using the true values of the explanatory variable, in other words, without measurement error, X_{ij}. The naive estimates were calculated using the explanatory variable with measurement error, Wij, but ignore this fact and include no attempt to fix this bias. The two SIMEX estimators, with linear and quadratic extrapolation techniques, were then used to attempt to fix this bias and returned their own respective estimates. All estimated model parameters were computed using the `lmer` function in R (Bates et al. 2014).

We investigated the effects of three different factors on the performance of these estimators. These factors were the reliability ratio, the number of level 2 units and the number of level 1 units. We describe each of these factors next.

Since the impact of measurement error is driven by its variance, the main focus of the simulation was to determine how effective the SIMEX process was across multiple levels of measurement error variance. In our simulation, we selected three levels of reliability ratios to apply to the modeling process. Reliability ratio is defined as

$$\kappa = \sigma_x^2/(\sigma_u^2 + \sigma_x^2)$$

where σ_u^2 is the measurement error variance and σ_x^2 is the variance of our true explanatory variable. Since our true explanatory variable, X, was constant throughout the simulation process, its variance remained constant and thus by changing the reliability ratio, we effectively changed the measurement error variance. Examining the formula above reveals that the reliability ratio and measurement error variance have an inverse relationship. This an increase in our ratio results in a decrease in our measurement error variance. The reliability ratios selected for the simulation were 0.75, 0.85, and 0.95.

The next two variables that were altered for the simulation were the number of level 2 units and the number of level 1 units. It is well-known that sample sizes can influence the results of a statistical analysis. Since a multilevel model has essentially two different sample sizes, both were deemed as important factors to vary across the simulation and observe how they affect the SIMEX estimates. The number of level 2 units in the simulation were 5 and 30. The number of level 1 units were 10 and 60.

For each combination of n, m, and κ, and using the model defined in (3), a dataset was created with n groups and m subjects within each group. Under each one of these combinations, 125 iterations of the simulation were completed, resulting in 125 sets of parameter estimates for each of the four estimation methods. For the two SIMEX estimators, we defined $\lambda = (0, 0.15, 0.25, 0.5, 0.75, 1)$. It should be noted that when $\lambda = 0$, we are simply referring to our observed variable, W, which was already used to fit the naive model. Thus in our actual SIMEX process, we only applied the other five values. The Monte Carlo step in equation (2) was completed using $B = 50$ Monte Carlo Replicates. The 50 replicates were averaged to give a single value for each estimate at each level of

λ. Using these mean values, two regression models, a linear and a quadratic, were fit for each estimated parameter to describe the trend of $\widehat{\theta}(\lambda)$ over λ. Each of these models were then extrapolated backwards to $\lambda = -1$ to give an estimated $\widehat{\theta}_{true}$ for each parameter. Thus, after every simulation iteration, we are left with 125 estimates from each of the four fitted models for each of the four parameters for each of the twelve factor combinations. These estimates were used to summarize and compare the distribution for each estimation method. This was accomplished by calculating the mean, bias, variance, and mean squared error. While we recognize that most simulations use a much larger number of iterations, this process was very computer intensive and constrained by the available technology. The next section describes the results of this simulation.

3.1 Results

The simulation provided some valuable insights in regards to the effectiveness of the SIMEX method. Our analysis of these results will focus on the influence that the three factors had on bias. Refer to Figure 7 for a full summary of the results, including variance of the estimates and mean square error. One major, and obvious, result from the simulation is how poorly SIMEX performed in estimating the random intercept variance, σ^2_{b0}. As Figure 7 reveals, along with Figures 2-4, the majority of the estimates for both the linear and quadratic approach returned negative values. With this flaw, the estimation of the random intercept variance will not be discussed in the analysis of the results, and instead we will save it for the conclusion and discuss possible improvements. In our discussion, we define a method as better if it produces a smaller bias.

The simulation suggests that the quadratic SIMEX is the most effective method for correcting measurement error bias. Since the quadratic fit appears to be at least as good as the linear fit, and usually better, the analysis will focus on the quadratic SIMEX method. First, let's examine the estimates strictly in terms of the different levels of reliability ratio. Figure 2 reveals the trends among mean estimates for each of the four parameters across the three different levels of reliability. From these plots, it is apparent that as reliability ratio decreases, and thus measurement error increases, the SIMEX method performs significantly better than the naive estimator. Welch's t-tests were performed to determine if the observed differences were statistically significant. When comparing the fixed slope estimates at the reliability ratio 0.95, none of the four methods produce significantly different results ($p > 0.05$). For a reliability of 0.85, the SIMEX quadratic method produces a significantly better result than the naive estimator ($p \approx 0$), while not being significantly different than the true estimator ($p = 0.50$). The quadratic SIMEX and true estimators were also found to not be significantly different at a reliability level 0.75 ($p = 0.13$). For the random slope and fixed intercept estimates, the tests show that the quadratic SIMEX method is significantly better than the naive model across all levels of reliability ratios ($p < 0.05$) and only outperformed by the true estimates at a reliability ratio of 0.75 ($p \approx 0$).

Analyzing the results with respect to number of level 2 units sampled shows similar trends between estimation methods, though it reveals that the number of units does not seem to have an effect. Figure 3 shows that as we increase from 5 to 30 units, our estimates tend to stay the same. Again, Welch's t-tests were used to test for significant differences between the methods. At a significance level 0.05, the SIMEX quadratic estimation method produced significantly better results than the naive estimator. There was also found to be no significant difference between the SIMEX quadratic and the true estimates for the fixed slope at both level 2 unit sample sizes, as well as not being different for the random slope estimates at a sample size of 5.

The final step in the analysis was to examine the effects of level 1 sample sizes on our estimates. Figure 4 shows that the number of level 1 samples did not have a large influence on the random and fixed slope estimates, but did significantly increase the bias in the fixed intercept estimates for the linear SIMEX and naive methods. Welch's t-tests returned similar results as the level 2 analysis, where at a significance level 0.05, the quadratic SIMEX method was significantly better than the naive method ($p \approx 0$). It was also observed that there was no significant difference between the quadratic SIMEX and true estimates for each parameter at a level 1 sample size of 10 and the fixed slope effect with sample size 60 ($p > 0.05$).

4 Data Analysis Example

The SIMEX method was used to analyze a dataset as an example for how one would apply it to real data. The selected data was from an orthodontics study which involved the measure of distance (mm) between the center of the pituitary to the pterygomaxillary fissure. The measurements were taken from 16 male children at ages 8, 10, 12, and 14 years old. Figure 5 provides a plot of the data, where each line represents a level 2 unit and each point is a level 1 unit. Specifically for this example, each child is a level 2 unit and each tooth measurement is a level 1 unit.

The measurement error introduced here is the result of rounding the age to an integer. Before the analysis, there are several assumptions that should be noted in order to use this dataset. The first is that despite this being a repeated measures study, we will just assume that the errors between level 1 units within a level 2 unit are independent. The second is that the ages are rounded to the closest birthday, and not just the current age class. For example, someone 3 months away from turning 10 will be considered 10 and not 9 as is the usual case. This allows us to then assume that the measurement errors have a mean of 0. We will also assume that these measurement errors, U, are normally distributed as defined in equation (1) The final assumption for this data is that the reliability ratio is 0.9. The observed variance for age is approximately 5, thus our estimated measurement error variance is 0.866.

The first step was to fit the naive model, ignoring the measurement error. The naive parameter estimates were $\widehat{\sigma}^2_{b0} = 7.11, \widehat{\sigma}^2_{b1} = 0.035, \widehat{\beta}_0 = 16.34$, and $\widehat{\beta}_1 = 0.78$. Next, we performed the SIMEX method, again using 50 repetitions for the simulation of each of our 5 λ values. Figure 6 shows the four SIMEX plots used to fit the models and create the SIMEX prediction, just as in Figure 1.

Since our simulation suggested that the quadratic fit provided the better estimates, we used that same method here. From the quadratic SIMEX model, the parameter estimates are $\widehat{\sigma}_{b0}^2 = 18.85, \widehat{\sigma}_{b1}^2 = -0.025, \widehat{\beta}_0 = 15.092$, and $\widehat{\beta}_1 = 1.328$. The important result here is that the estimate of the fixed slope has increased in value from the naive estimate. Knowing that measurement error causes an attenuation bias, the fact that the SIMEX method has increased the value of the estimated slope suggests that SIMEX was successful in correcting for the measurement error bias. Also of interest is that in this example it was the random slope variance that predicted a negative value, in contrast to the simulation producing negative estimates for the random intercept variance. One possible explanation is that random slope variance is very close to zero, suggesting that there is very little variation of slopes among level 2 units. This possibly lead to the poor estimation results from the SIMEX process. Based on our simulation and contingent on our assumption, we claim that these SIMEX estimates provide less biased results than the naive estimates and thus a better understanding of the relationship between age and distance.

5 Conclusion

Through this study, we were able to evaluate the effectiveness of the SIMEX method for correcting bias in a multilevel model. Based on the results of the simulation, it appears that SIMEX is a valid and useful tool for establishing statistically sound parameter estimates. Though, as seen in the appearance of negative variance estimates in both our simulation and data analysis example, further research is needed to enhance the way SIMEX creates these estimates to prevent such errors from occurring. Future research should also look to expand on the multilevel models being studied, such as those with more than one covariate or repeated measures with correlated errors.

6 Acknowledgments

I would like to thank my advisor, Dr. Julie McIntyre, for all the knowledge and assistance provided during this project, as well as being an integral part of my graduate education. I also would like to thank Dr. Ron Barry, Dr. Lubov Zeifman, and Dr. Margaret Short for their role in the completion of this project and in my outstanding education obtained at the University of Alaska.

7 References

Bates D, Maechler M, Bolker BM and Walker S (2014). lme4: Linear mixed-effects models using Eigen and S4. ArXiv e-print; submitted to Journal of Statistical Software, URL:http://arxiv.org/abs/1406.5823.

Cook, J., and Stefanski, L. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association, 1314-1314.

Goldstein, H. (1995). Chapter 2: The basic linear multilevel model and its estimation. In Multilevel statistical models (2nd ed.). London: E. Arnold.

Goldstein, H. (1995). Chapter 10: Multilevel models with measurement error. In Multilevel statistical models (2nd ed.). London: E. Arnold.

Fuller, W. (1987). Measurement error models. New York: Wiley.

8 Figures

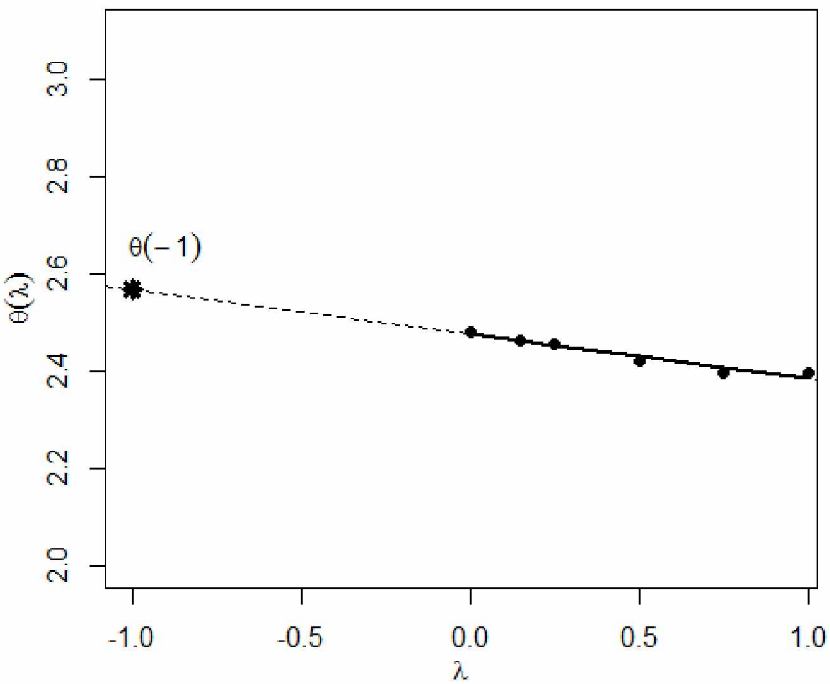

Figure 1: SIMEX method using a linear extrapolant

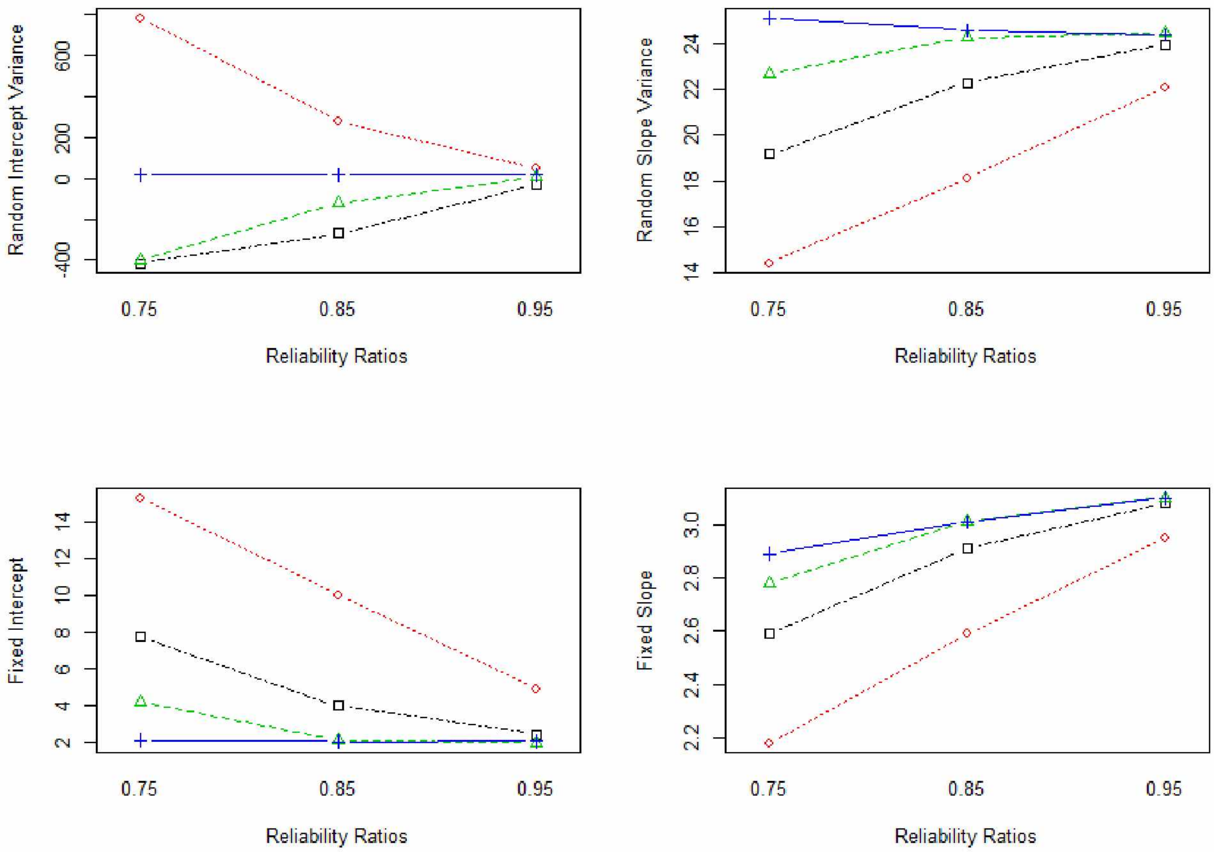

Figure 2: Parameter estimates for each method across each level of reliability ratio.

+ true estimate

△ quadratic SIMEX

□ linear SIMEX

◯ naive estimate

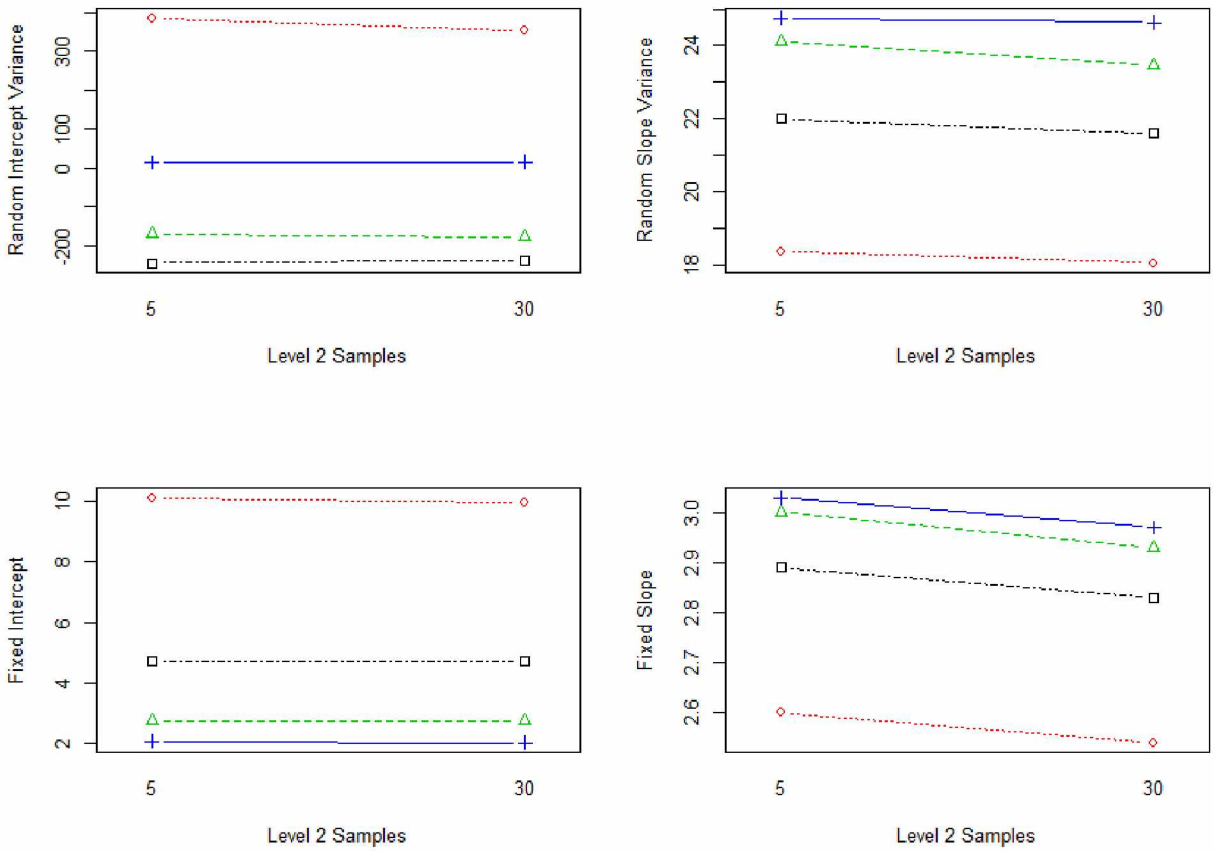

Figure 3: Parameter estimates for each method across each Level 2 sample size.

+ true estimate

△ quadratic SIMEX

□ linear SIMEX

◯ naive estimate

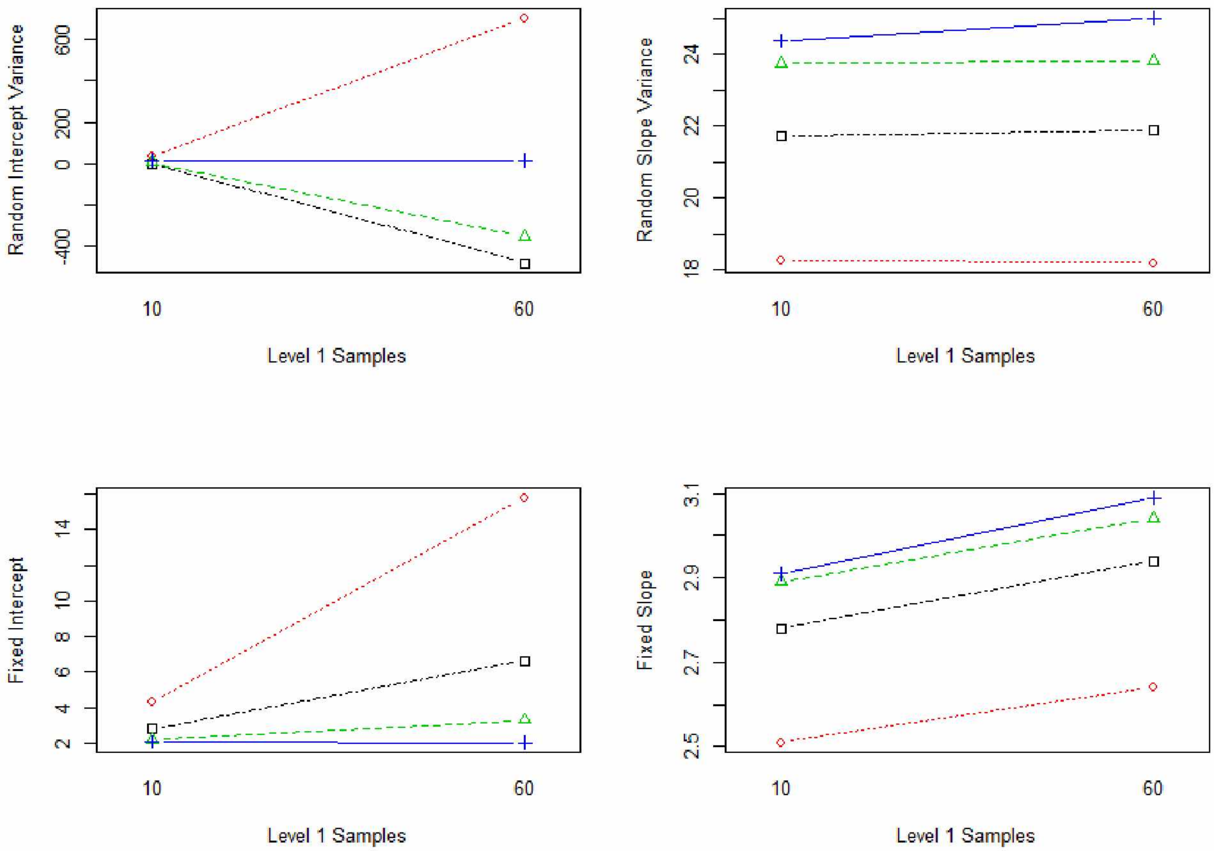

Figure 4: Parameter estimates for each method across each Level 1 sample size.

+ true estimate

△ quadratic SIMEX

□ linear SIMEX

◯ naive estimate

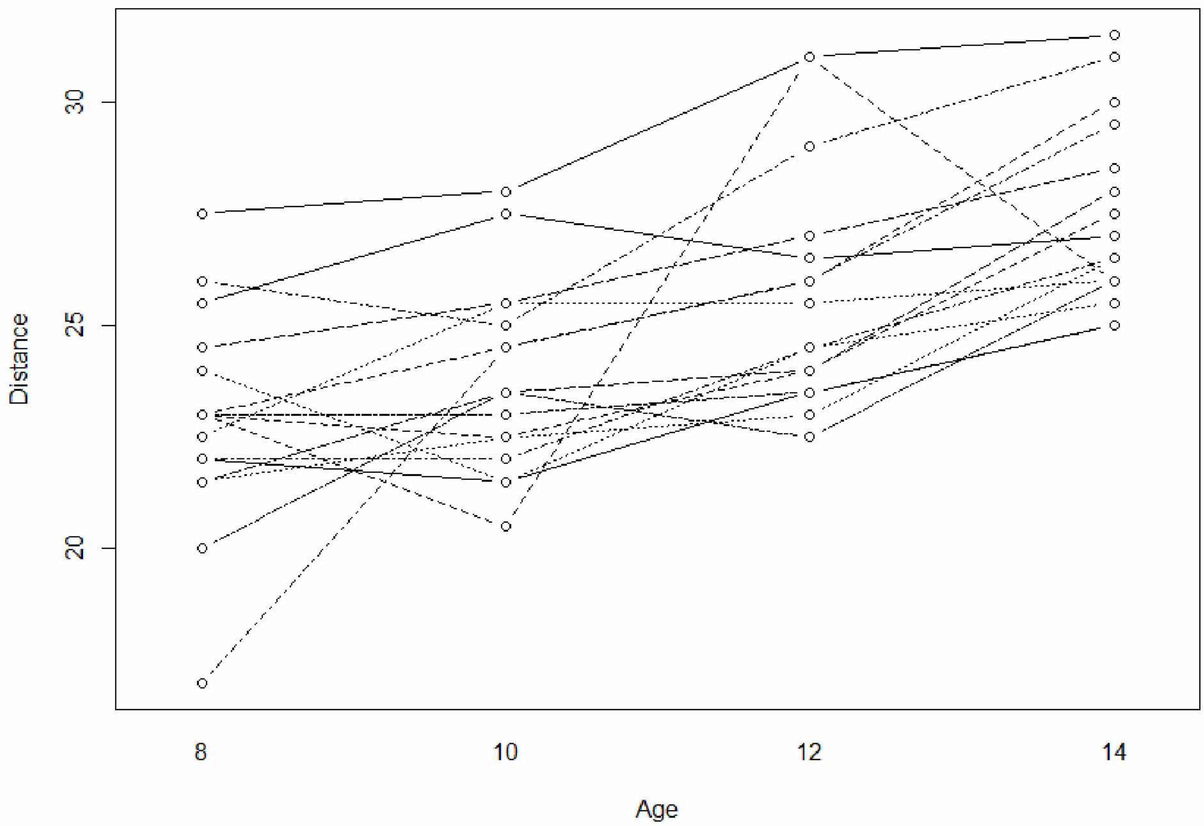

Figure 5: Orthodontist data

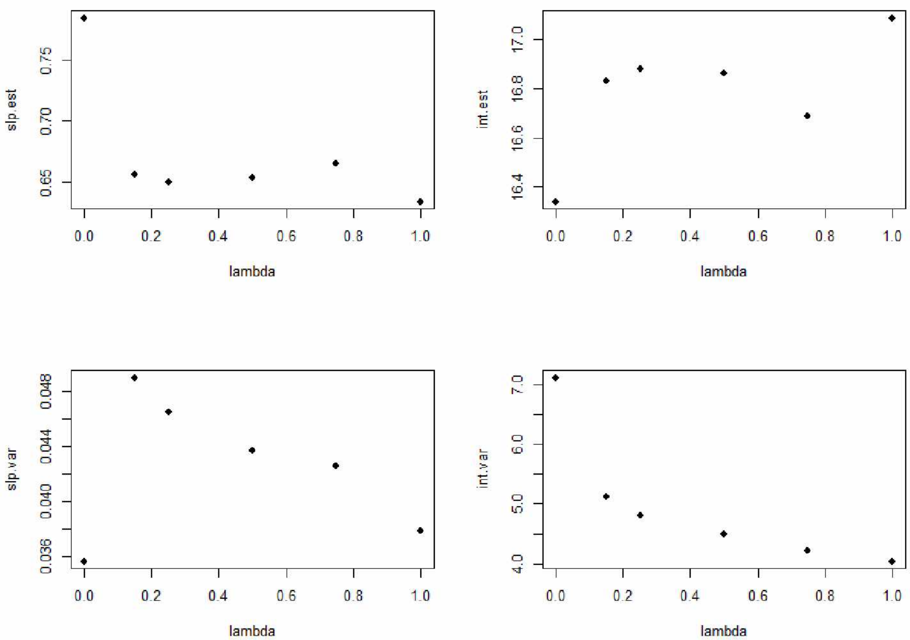

Figure 6: SIMEX extrapolation step for each method in orthodontics study

Random Intercept (b0=16)			TRUE			Naïve			Simex.Linear			Simex.Quadratic		
N.Group	N.Sample	RR	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance
5	10	0.95	14.51	-1.49	148.31	17.89	1.89	234.70	12.35	-3.65	353.48	17.25	1.25	617.31
		0.85	16.73	0.73	186.65	36.42	20.42	1380.54	6.31	-9.69	1423.41	9.47	-6.53	4099.33
		0.75	17.40	1.40	201.49	55.96	39.96	2694.94	-7.94	-23.94	4048.77	-7.54	-23.54	11759.22
	60	0.95	15.03	-0.97	127.09	83.13	67.13	18557.68	-71.46	-87.46	6741.38	-23.10	-39.10	30980.53
		0.85	15.06	-0.94	107.70	517.83	501.83	135778.54	-523.02	-539.02	169128.27	-246.31	-262.31	229773.34
		0.75	16.34	0.34	123.26	1606.83	1590.83	1737516.06	-878.85	-894.85	769501.22	-764.61	-780.61	1184971.46
30	10	0.95	16.49	0.49	26.03	17.29	1.29	52.52	15.85	-0.15	91.83	18.76	2.76	167.07
		0.85	16.17	0.17	23.42	26.39	10.39	173.20	-7.88	-23.88	243.77	8.48	-7.52	1163.20
		0.75	14.98	-1.02	23.91	55.38	39.38	688.80	-17.56	-33.56	635.42	-17.06	-33.06	3218.05
	60	0.95	16.03	0.03	21.43	73.53	57.53	1076.79	-91.48	-107.48	1483.39	11.11	-4.89	10216.05
		0.85	16.32	0.32	13.53	530.35	514.35	42583.45	-560.57	-576.57	40311.84	-273.29	-289.29	154022.76
		0.75	16.08	0.08	18.10	1412.09	1396.09	183451.82	-756.87	-772.87	70937.47	-805.24	-821.24	383572.00

Random Slope (b1=25)			TRUE			Naïve			Simex.Linear			Simex.Quadratic		
N.Group	N.Sample	RR	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance
5	10	0.95	25.12	0.12	322.75	22.88	-2.12	270.79	24.59	-0.41	314.37	25.10	0.10	364.68
		0.85	24.03	-0.97	290.00	18.43	-6.57	189.24	22.52	-2.48	296.57	24.99	-0.01	396.31
		0.75	23.83	-1.17	304.20	14.38	-10.62	134.25	19.04	-5.96	241.75	22.67	-2.33	429.43
	60	0.95	23.43	-1.57	338.67	21.02	-3.98	263.50	22.89	-2.11	310.43	23.34	-1.66	329.85
		0.85	24.61	-0.39	213.11	17.92	-7.08	115.04	22.12	-2.88	176.19	23.97	-1.03	207.79
		0.75	27.56	2.56	442.01	15.57	-9.43	142.80	20.83	-4.17	256.01	24.55	-0.45	355.19
30	10	0.95	23.79	-1.21	40.63	21.75	-3.25	32.67	23.50	-1.50	37.19	24.26	-0.74	54.78
		0.85	24.43	-0.57	46.24	17.90	-7.10	27.31	22.00	-3.00	44.01	23.64	-1.36	57.70
		0.75	25.06	0.06	35.78	14.19	-10.81	13.17	18.75	-6.25	24.55	21.89	-3.11	39.82
	60	0.95	25.20	0.20	32.92	22.77	-2.23	26.76	24.83	-0.17	31.69	25.11	0.11	32.89
		0.85	25.30	0.30	54.47	18.27	-6.73	27.77	22.53	-2.47	42.57	24.42	-0.58	51.27
		0.75	24.04	-0.96	33.65	13.52	-11.48	10.47	18.09	-6.91	18.63	21.59	-3.41	27.86

Fixed Intercept (B0=2)			TRUE			Naïve			Simex.Linear			Simex.Quadratic		
N.Group	N.Sample	RR	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance
5	10	0.95	1.91	-0.09	4.13	2.71	0.71	4.83	2.08	0.08	5.14	2.08	0.08	5.59
		0.85	2.06	0.06	4.38	4.28	2.28	10.43	2.59	0.59	10.24	1.64	-0.36	14.77
		0.75	2.34	0.34	3.87	5.86	3.86	14.24	3.70	1.70	12.43	2.60	0.60	23.73
	60	0.95	2.18	0.18	3.22	7.70	5.70	21.40	3.08	1.08	12.17	2.02	0.02	21.17
		0.85	1.99	-0.01	3.03	15.41	13.41	133.03	4.98	2.98	30.91	1.79	-0.21	43.88
		0.75	1.92	-0.08	2.79	24.86	22.86	228.34	11.86	9.86	71.65	6.26	4.26	88.91
30	10	0.95	1.99	-0.01	0.73	2.74	0.74	1.08	2.10	0.10	1.05	1.88	-0.12	2.65
		0.85	2.03	0.03	0.49	4.29	2.29	1.62	2.57	0.57	1.45	2.17	0.17	3.30
		0.75	2.01	0.01	0.55	5.92	3.92	2.35	3.72	1.72	1.73	2.67	0.67	3.76
	60	0.95	1.99	-0.01	0.52	6.38	4.38	3.53	2.37	0.37	2.37	1.76	-0.24	7.43
		0.85	1.96	-0.04	0.47	15.93	13.93	24.59	5.71	3.71	7.23	2.74	0.74	18.08
		0.75	2.05	0.05	0.67	24.53	22.53	49.87	11.67	9.67	15.17	5.20	3.20	17.76

Fixed Slope (B1=3)			TRUE			Naïve			Simex.Linear			Simex.Quadratic		
N.Group	N.Sample	RR	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance	Estimate	Bias	Variance
5	10	0.95	2.94	-0.06	4.33	2.80	-0.20	3.94	2.91	-0.09	4.27	2.92	-0.08	4.38
		0.85	2.98	-0.02	4.36	2.58	-0.42	3.43	2.89	-0.11	4.35	3.05	0.05	5.09
		0.75	2.73	-0.27	4.22	2.07	-0.93	2.75	2.46	-0.54	3.98	2.66	-0.34	4.95
	60	0.95	3.44	0.44	4.76	3.26	0.26	4.28	3.41	0.41	4.68	3.44	0.44	4.81
		0.85	3.12	0.12	5.91	2.67	-0.33	4.30	3.01	0.01	5.48	3.11	0.11	5.79
		0.75	2.98	-0.02	3.72	2.22	-0.78	2.13	2.65	-0.35	3.03	2.83	-0.17	3.53
30	10	0.95	3.06	0.06	1.07	2.93	-0.07	0.97	3.05	0.05	1.05	3.08	0.08	1.14
		0.85	2.87	-0.13	0.97	2.47	-0.53	0.74	2.77	-0.23	0.95	2.84	-0.16	1.06
		0.75	2.91	-0.09	0.81	2.21	-0.79	0.47	2.61	-0.39	0.65	2.79	-0.21	0.83
	60	0.95	2.96	-0.04	0.83	2.81	-0.19	0.75	2.95	-0.05	0.82	2.96	-0.04	0.84
		0.85	3.09	0.09	1.04	2.63	-0.37	0.75	2.97	-0.03	0.96	3.06	0.06	1.03
		0.75	2.94	-0.06	0.83	2.21	-0.79	0.47	2.63	-0.37	0.66	2.84	-0.16	0.77

Figure 7: Summary results from full simulation